

# Minería de datos y Patrones

Version 2025-I

## Presentación

Dr. José Ramón Iglesias  
DSP-ASIC BUILDER GROUP  
Director Semillero TRIAC  
Ingeniería Electrónica  
Universidad Popular del Cesar

# **Introducción**

## Definición

La minería de datos tiene como objetivo la **extracción** de conocimiento **a partir de grandes cantidades de datos**, mediante métodos automáticos o semiautomáticos.

Propone usar un conjunto de **algoritmos** [...] para construir modelos a partir de los datos, es decir, encontrar estructuras interesantes o patrones según criterios predefinidos, y **extraer un máximo de conocimientos** de ellos.<sup>1</sup>

---

<sup>1</sup>[https://fr.wikipedia.org/wiki/Exploration\\_de\\_donn'ees](https://fr.wikipedia.org/wiki/Exploration_de_donn%C3%A9es)

## Diferentes tipos de datos:

- Datos estructurados:
  - Datos sociales: Edad, Salario, Color de piel, Lugar de residencia
  - Datos métricos: *Likes* de una publicación, Tiempo pasado en una página, Número de amigos en común
- Datos no estructurados:
  - Texto: Frase, Párrafo, Documento
  - Sonido: Canción, Discurso
  - Imagen: Foto, Vídeo

## Diferentes tipos de Minería:

- Exploración de datos: Detectar valores simples, sesgos
- Tarea de clasificación/regresión: Alimentarse de datos para caracterizar nuevos datos **por clase o con un valor**, de manera supervisada
- Tarea de agrupamiento: Caracterizar datos por clase de manera no supervisada
- Reducción de dimensiones: Desarrollar estructuras comunes para representaciones comprimidas de datos

[Aplicaciones](#)

[Significacion de las termas](#)

[Prerrequisitos](#)

[TP Exploración de Datos:](#)

[MovieLens](#)

[Overview](#)

- Detección de eventos en un texto



- Detección de eventos en un texto



- Procesamiento automático de opiniones de usuarios





- Detección de eventos en un texto



- Procesamiento automático de opiniones de usuarios



- Propuesta de recomendaciones a un usuario

Les connaissez-vous ?



**Marina Dunion**  
Digital Marketing @Air France  
& Co-Founder @FlexFly  
📞 Teddy Viraye-  
Chevalier et 3 autres  
relations

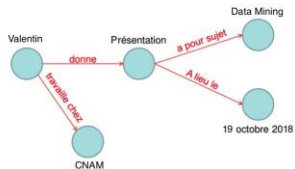


**Salvatore Anzalone**  
Post-Doc at ISIR, University  
Pierre et Marie Curie, Paris  
📞 Thomas Janssoone et  
2 autres relations

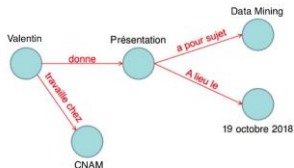


**Halla Olafsdottir**  
Medical Solutions Project  
Manager | Chef de Projet  
📞 Télécom ParisTech

- Detección de relaciones entre entidades en un texto



- Detección de relaciones entre entidades en un texto



- Respuesta a una pregunta

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

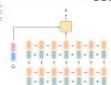


Document  
Retriever

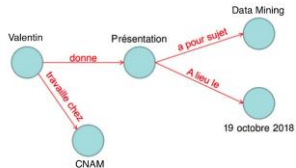


Document  
Reader

833,500



- Detección de relaciones entre entidades en un texto
- Respuesta a una pregunta
- Módulo de IE para un agente conversacional



Q: How many of Warsaw's inhabitants spoke Polish in 1933?



# Outline : Significacion de las termas

[Aplicaciones](#)

[Significacion de las termas](#)

[Prerrequisitos](#)

[TP Exploración de Datos:](#)

[MovieLens](#)

[Overview](#)

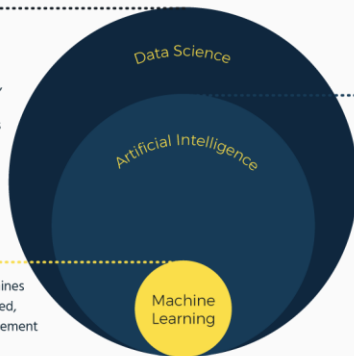
## AI vs. Data Science vs. Machine Learning

### Data Science

- Collection, preparation, and analysis of data
- Leverages AI/ML, research, industry expertise, and statistics to make business decisions

### Machine Learning

- Algorithms that help machines improve through supervised, unsupervised, and reinforcement learning
- Subset of AI and Data Science tool



### Artificial Intelligence

- Technology for machines to understand/interpret, learn, and make 'intelligent' decisions
- Includes Machine Learning among many other fields

**Figure 1:** Diferencias entre campos

## En resumen

Data Science se centra en el análisis de datos para extraer conocimiento, Machine Learning utiliza algoritmos para hacer predicciones y tomar decisiones basadas en datos, y Artificial Intelligence se refiere al desarrollo de sistemas que pueden realizar tareas inteligentes de manera autónoma.

Definición (sobre) simplista:

- Data mining genera entendimiento.
- Machine learning genera predicciones.
- Artificial intelligence genera acciones.

# Ejemplo en plataforma de musica

## **Data Scientist**

Recopila y analiza datos de usuarios de plataformas de música para identificar patrones y preferencias musicales.

## **Machine Learner**

Desarrolla y optimiza un modelo de recomendación de música utilizando algoritmos de aprendizaje automático para predecir las preferencias de los usuarios.

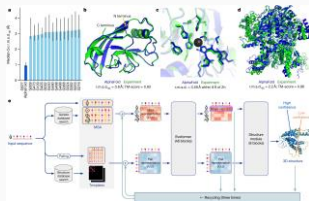
## **Artificial Intelligence**

Implementa un agente social que puede interactuar con el usor, para mejorar la personalización de las recomendaciones musicales y proporcionar una experiencia más precisa y contextualizada.



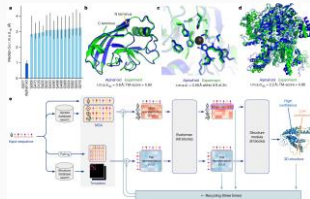
# Significacion del trabajo: Porque hacer eso?

- Avance científico



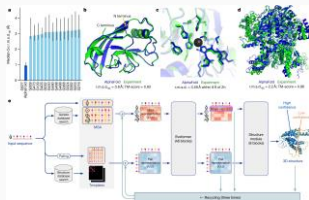
# Significación del trabajo: Porque hacer eso?

- Avance científico
- Prevención y gestión de desastres naturales



# Significación del trabajo: Porque hacer eso?

- Avance científico
- Prevención y gestión de desastres naturales
- Impacto en la salud pública



## Open Chronic

Améliorer la prise en charge des malades chroniques

Santé Promotion 3

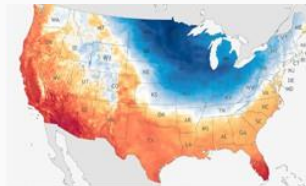
Ministère de la santé, Direction de la recherche, des études, d et des statistiques

Paris

Data science

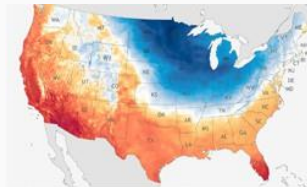
# Significacion del trabajo: Porque hacer eso?

- Sostenibilidad ambiental

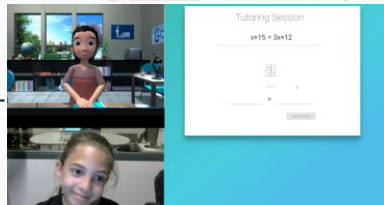


# Significación del trabajo: Porque hacer eso?

- Sostenibilidad ambiental

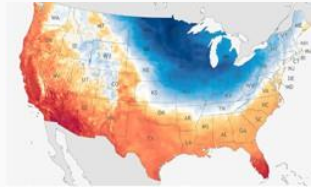


- Impulso a la educación y la investigación

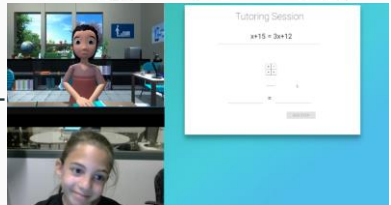


# Significación del trabajo: Porque hacer eso?

- Sostenibilidad ambiental



- Impulso a la educación y la investigación



- Democracia participativa



[Aplicaciones](#)

[Significacion de las temas](#)

[Prerrequisitos](#)

[TP Exploración de Datos:](#)

[MovieLens](#)

[Overview](#)

## Partes teóricas

Bases de estadística, álgebra lineal: Presentación general del aprendizaje estadístico, Bases matemáticas de los diferentes modelos, enfoque intuitivo

## Partes prácticas

Bases de Python: Uso de herramientas de manipulación de datos, Uso de una biblioteca de DL, Uso de una biblioteca de ML, Análisis de sentimientos, Ranking sobre preferencias de vino, ...<sup>2</sup>

---

<sup>2</sup>Non contractual por este curso



# Material

- Computadora
- Jupyter Notebook y Anaconda:  
<https://www.anaconda.com/download/>
- Los notebooks y las cheatsheets disponibles online:

### Python For Data Science Cheat Sheet

#### NumPy Basics

Learn Python for Data Science interactively at: [www.DataCamp.com](http://www.DataCamp.com)

**NumPy**

The NumPy library is the core library for scientific computing in Python. It provides a high-performance multidimensional array object, and tools for working with these arrays.

Use the following import convention:

```
>>> import numpy as np
```

**NumPy Arrays**

**1D array**

```
>>> np.array([1, 2, 3])
```

**2D array**

```
>>> np.array([[1, 2, 3], [4, 5, 6]])
```

**3D array**

```
>>> np.array([[[1, 2, 3], [4, 5, 6]], [[7, 8, 9], [10, 11, 12]]])
```

**Creating Arrays**

```
>>> np.zeros(10)
```

### Python For Data Science Cheat Sheet

#### Matplotlib

Learn Python for Data Science interactively at: [www.DataCamp.com](http://www.DataCamp.com)

**Matplotlib**

Matplotlib is a Python 2D plotting library which produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms.

**1 Prepare The Data**

```
>>> import numpy as np
>>> x = np.linspace(0, 10, 100)
>>> y = np.sin(x)
>>> X = np.vstack([x, y])
```

**2D Data or Image**

```
>>> data = 2 * np.random.random((10, 10))
>>> data2 = 3 * np.random.random((10, 10))
>>> X = np.vstack([data, data2])
>>> X = X * 255
```

### Python For Data Science Cheat Sheet

#### Scikit-Learn

Learn Python for data science interactively at: [www.DataCamp.com](http://www.DataCamp.com)

**Scikit-learn**

Scikit-learn is an open source Python library that implements a range of machine learning, preprocessing, cross-validation and visualization algorithms using a unified interface.

**A Basic Example**

```
>>> from sklearn import datasets, models, preprocessing
>>> from sklearn.model_selection import train_test_split
>>> data = datasets.load_digits()
>>> X, y = data.data[:, :, :], data.target
>>> X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0)
>>> model = preprocessing.StandardScaler().fit(X_train)
>>> X_train = model.transform(X_train)
>>> X_test = model.transform(X_test)
>>> model = models.LinearSVC().fit(X_train, y_train)
>>> y_pred = model.predict(X_test)
>>> accuracy_score(y_test, y_pred)
```

### Python For Data Science Cheat Sheet

#### Pandas Basics

Learn Python for Data Science interactively at: [www.DataCamp.com](http://www.DataCamp.com)

**Pandas**

The Pandas library is built on NumPy and provides easy-to-use data structures and data analysis tools for the Python programming language.

**pandas**

Use the following import convention:

```
>>> import pandas as pd
```

**Pandas Data Structures**

**Series**

A one-dimensional labeled array capable of holding any data type

```
>>> pd.Series([1, 2, 3, 4])
```

### Python For Data Science Cheat Sheet

#### Jupyter Notebook

Learn More Python for Data Science interactively at: [www.DataCamp.com](http://www.DataCamp.com)

**Saving/Loading Notebooks**

Create new notebook

Make a copy of the current notebook

Save current notebook and recent checkpoint

Preview of the printed notebook

Close notebook & stop running any scripts

Writing Code And Text

### Python For Data Science Cheat Sheet

#### Keras

Learn Python for data science interactively at: [www.DataCamp.com](http://www.DataCamp.com)

**Keras**

Keras is a powerful and easy-to-use deep learning library for Theano and TensorFlow that provides a high-level neural networks API to develop and evaluate deep learning models.

**A Basic Example**

```
>>> import numpy as np
>>> from keras.models import Sequential
>>> from keras.layers import Dense
>>> data = np.random.random((1000, 100))
>>> labels = np.random.random((1000, 1))
>>> model = Sequential()
>>> model.add(Dense(100, activation='tanh'))
>>> model.add(Dense(1, activation='sigmoid'))
>>> model.compile(optimizer='rmsprop',
>>>               loss='binary_crossentropy',
>>>               metrics=['accuracy'])
>>> model.fit(data, labels, epochs=10, batch_size=32)
```

### Python For Data Science Cheat Sheet

#### Supervised Learning

Learn Python for data science interactively at: [www.DataCamp.com](http://www.DataCamp.com)

**Supervised Learning**

Supervised learning is a type of machine learning where the model is trained on a dataset with known labels. The goal is to learn a function that maps input features to output labels.

**Linear Regression**

Linear regression is a supervised learning algorithm that models the relationship between a continuous target variable and one or more input features.

**Logistic Regression**

Logistic regression is a supervised learning algorithm that models the probability of a binary outcome given a set of input features.

**Support Vector Machines (SVM)**

Support vector machines are supervised learning models that find the optimal hyperplane that separates the data into two classes.

**Decision Trees**

Decision trees are supervised learning models that use a series of binary decisions to partition the data into regions and assign a class label to each region.

**Random Forests**

Random forests are supervised learning models that combine the predictions of many individual decision trees to improve the overall performance.

**Gradient Boosting**

Gradient boosting is a supervised learning technique that builds an ensemble of weak models, where each model is trained to correct the errors of the previous model.

### Python For Data Science Cheat Sheet

#### Model Evaluation

Learn Python for data science interactively at: [www.DataCamp.com](http://www.DataCamp.com)

**Model Evaluation**

Model evaluation is the process of assessing the performance of a machine learning model on new, unseen data. This is typically done using metrics such as accuracy, precision, recall, and F1 score.

**Cross-Validation**

Cross-validation is a technique for evaluating the performance of a model by partitioning the data into training and testing sets multiple times and averaging the results.

**Confusion Matrix**

A confusion matrix is a table that compares the predicted and actual outcomes of a classification model. It provides a detailed view of the model's performance, including true positives, false positives, true negatives, and false negatives.

**Precision-Recall Curve**

A precision-recall curve is a plot that shows the relationship between precision and recall for a classification model. It is useful for comparing different models and understanding the trade-off between precision and recall.

**ROC Curve**

A receiver operating characteristic (ROC) curve is a plot that shows the relationship between the true positive rate and the false positive rate for a classification model. It is a common way to evaluate the performance of binary classifiers.

# Outline : TP Exploración de Datos: MovieLens

[Aplicaciones](#)

[Significación de las temáticas](#)

[Prerrequisitos](#)

[TP Exploración de Datos:  
MovieLens](#)

[Overview](#)

Estudio simple de un conjunto de datos de críticas de películas

- 3 millones de notas
- Descriptores sociales: edad, sexo, ...
- Primer enfoque básico de minería de datos

The logo for MovieLens, featuring the word "movielens" in a lowercase, orange, sans-serif font.

Pandas



- Biblioteca de Python para manipular bases de datos:  
<https://pandas.pydata.org/>
- Permite realizar operaciones y visualizaciones
- Fácil de usar

[Aplicaciones](#)

[Significacion de las termas](#)

[Prerrequisitos](#)

[TP Exploración de Datos:](#)

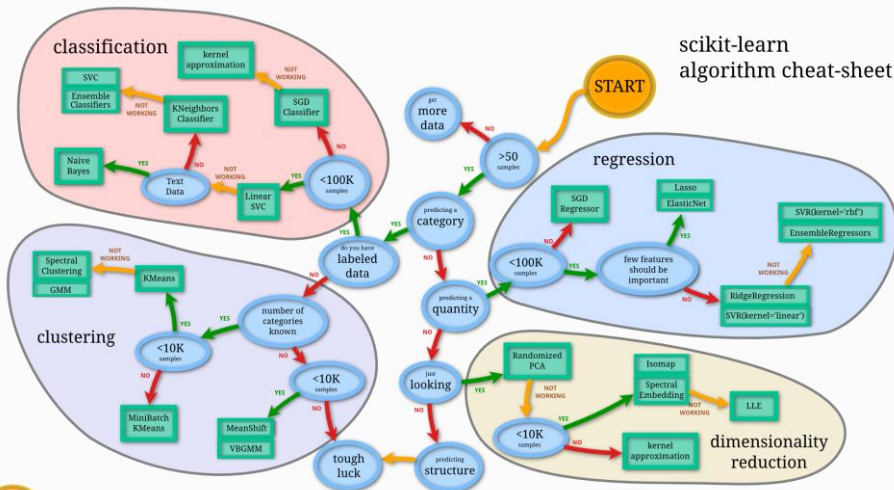
[MovieLens](#)

[Overview](#)

- Clasificación: predecir una clase determinada
- Regresión: predecir un valor
- Clustering: agrupar elementos en cluster (clases no determinadas)
- Reducción de dimensión: disminuir el espacio de representación de los datos
- Detección de anomalía

# Los diferentes métodos

## scikit-learn algorithm cheat-sheet



Back

## Ejemplos: Classificacion

- Reconocimiento de emociones en el habla:
- Clasificación de especies de animales a partir de imágenes:
- Detección de objetos en imágenes médicas:



## Ejemplos: Classificacion

- Reconocimiento de emociones en el habla: **la persona esta enojada o feliz**
- Clasificación de especies de animales a partir de imágenes:
- Detección de objetos en imágenes médicas:

## Ejemplos: Classificacion

- Reconocimiento de emociones en el habla: **la persona esta enojada o feliz**
- Clasificación de especies de animales a partir de imágenes: **es un gato o un puma?**
- Detección de objetos en imágenes médicas:

## Ejemplos: Classificacion

- Reconocimiento de emociones en el habla: **la persona esta enojada o feliz**
- Clasificación de especies de animales a partir de imágenes: **es un gato o un puma?**
- Detección de objetos en imágenes médicas: **es un tumor?**

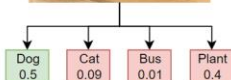
# Ejemplos: Clasificación

- Reconocimiento de emociones en el habla: **la persona esta enojada o feliz**
- Clasificación de especies de animales a partir de imágenes: **es un gato o un puma?**
- Detección de objetos en imágenes médicas: **es un tumor?**

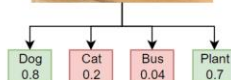
Binary Classification



Multiclass Classification



Multilabel Classification



## Ejemplos: Regresion

- Reconocimiento de emociones en el habla:
- Evaluación de daños a partir de imágenes despues un terremoto:
- Detección de severidad de Alzheimer en la voz:

## Ejemplos: Regresion

- Reconocimiento de emociones en el habla: **cual es la intesidad del enojo?**
- Evaluación de daños a partir de imágenes despues un terremoto:
- Detección de severidad de Alzheimer en la voz:

## Ejemplos: Regresion

- Reconocimiento de emociones en el habla: **cual es la intesidad del enojo?**
- Evaluación de daños a partir de imágenes despues un terremoto: **la ambulancia puede utilizar el puente?**
- Detección de severidad de Alzheimer en la voz:

## Ejemplos: Regresion

- Reconocimiento de emociones en el habla: **cual es la intesidad del enojo?**
- Evaluación de daños a partir de imágenes despues un terremoto: **la ambulancia puede utilizar el puente?**
- Detección de severidad de Alzheimer en la voz: **como avanzado es el estado?**



# Ejemplos: Regresion

- Reconocimiento de emociones en el habla: **cual es la intensidad del enojo?**
- Evaluación de daños a partir de imágenes despues un terremoto: **la ambulancia puede utilizar el puente?**
- Detección de severidad de Alzheimer en la voz: **como avanzado es el estado?**

## Age Prediction via Regression



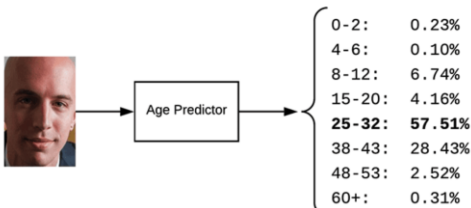
# Ejemplos: Regresion

- Reconocimiento de emociones en el habla: **cual es la intensidad del enojo?**
- Evaluación de daños a partir de imágenes despues un terremoto: **la ambulancia puede utilizar el puente?**
- Detección de severidad de Alzheimer en la voz: **como avanzado es el estado?**

## Age Prediction via Regression



## Age Prediction via Classification



## Ejemplos: Clustering

- Topic Mining en forums politicos:
- Clasificación de desinformación en redes sociales:
- Segmentación de clientes:

## Ejemplos: Clustering

- Topic Mining en forums politicos: **de que los ciudadanos se preocupan?**
- Clasificación de desinformación en redes sociales:
- Segmentación de clientes:

## Ejemplos: Clustering

- Topic Mining en forums politicos: **de que los ciudadanos se preocupan?**
- Clasificación de desinformación en redes sociales: **a grupamos estas noticias que parecen raras**
- Segmentación de clientes:

## Ejemplos: Clustering

- Topic Mining en forums politicos: **de que los ciudadanos se preocupan?**
- Clasificación de desinformación en redes sociales: **a grupamos estas noticias que parecen raras**
- Segmentación de clientes: **la gente que le gustan las chelas**

# Ejemplos: Clustering

- Topic Mining en forums politicos: **de que los ciudadanos se preocupan?**
- Clasificación de desinformación en redes sociales: **agrupamos estas noticias que parecen raras**
- Segmentación de clientes: **la gente que le gustan las chelas**



**Questions?**